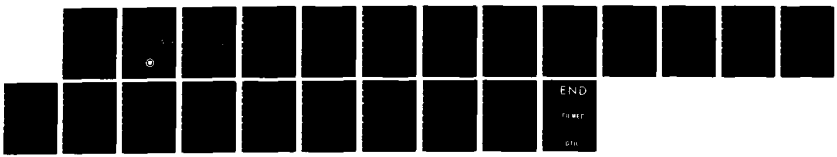
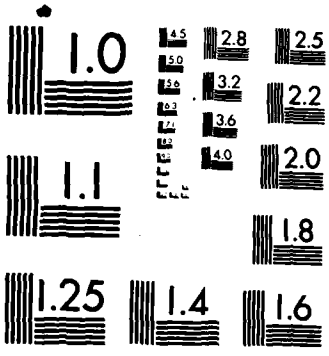
AD-A162 869 EFFECT OF ADDITIONAL VARIABLES IN PRINCIPAL COMPONENT ANALYSIS DISCRIMINA (U) PITTSBURGH UNIV PA CENTER FOR MULTIVARIATE ANALYSIS Y FUJIKOSHI ET AL AUG 85 1/1

UNCLASSIFIED TR-85-31 AFOSR-TR-85-0980 F49620-85-C-0008 F/G 12/1 NL

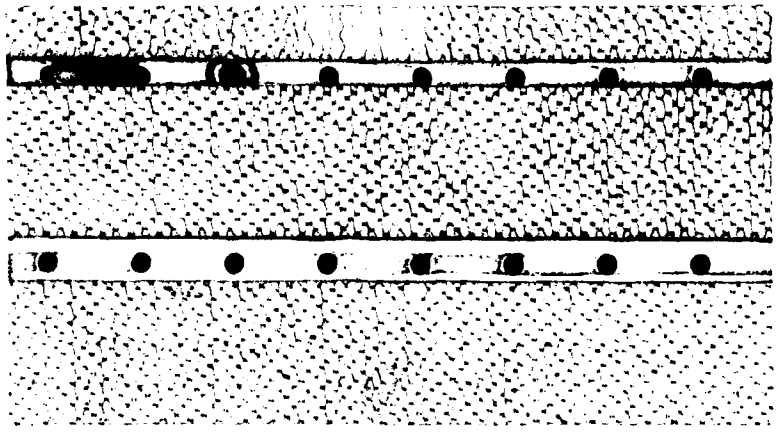




AFOSR-TR- 85-0980

(2)

AD-A162 069



DTIC
ELECTE
DEC 9 1985
S D
B

Center for Multivariate Analysis

University of Pittsburgh

DTIC FILE COPY

85 12 6 042

EFFECT OF ADDITIONAL VARIABLES IN PRINCIPAL COMPONENT ANALYSIS, DISCRIMINANT ANALYSIS AND CANONICAL CORRELATION ANALYSIS*

Y. Fujikoshi
Hiroshima University

P. R. Krishnaiah
University of Pittsburgh

J. Schmidhammer
University of Tennessee

August 1985

Technical Report 85-31

Center for Multivariate Analysis
University of Pittsburgh
515 Thackeray Hall
Pittsburgh, PA 15260

* Research sponsored by the Air Force Office of Scientific Research (AFSC), under contracts F49620-82-K-0001 and F49620-85-C-0008. This work was performed at the Center for Multivariate Analysis. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

1. INTRODUCTION

In a number of situations, it is of interest to find out whether the addition of a new set of variables gives additional information for inference. For example, in the area of principal component analysis, it is of interest to find out whether the new variables contribute to explanation of the variation among experimental units. In the area of multi-group discriminant ananlysis, it is important to find out whether the addition of new variables contributes to the discrimination between the groups. Similarly, in the area of canonical correlation analysis, it is of interest to find out as to whether the addition of variables to one or both sets of variables contributes to the degree of association between the two sets of variables.

Rao (1966) considered the effect of additional variables on the efficiency of estimates and the power of the test under multivariate regression model. Recently, Wijsman (1984) derived asymptotic distribution of the increase in the largest sample canonical correlation whem some variables are added. In Section 3 of this paper, we first derive asymptotic distributions of changes in functions of the eigenvalues of the sample covariance matrix. Asymptotic distributions of changes in functions of the eigenvalues of the multivariate analysis of variance (MANOVA) matrix when some variables are added are derived in Section 4. In Section 5, we derive asymptotic distributions of changes in certain functions of the sample canonical correlations when new variables are added to one or both sets of original variables. The above results are derived under the assumption that the underlying distribution is multivariate normal. Further results are given in Section 6.

2. PRELIMINARIES

In this section, we state the three lemmas which are needed in the sequel.

Lemma 2.1. [Cramér (1946, p.366)]. Let $\underset{\sim}{x}_n$ be a p-component random vector and $\underset{\sim}{\mu}$: $p \times 1$ be a fixed vector. Assume $\sqrt{n}\ (\underset{\sim}{x}_n - \underset{\sim}{u})$ converges to $N(\underset{\sim}{0},\Sigma)$ in law. Let $f(\underset{\sim}{x})$ be continuously differentiable in a neighborhood of $\underset{\sim}{\mu}$ and let $\underset{\sim}{\xi} = \partial f(\underset{\sim}{x})/\partial \underset{\sim}{x}\big|_{\underset{\sim}{x}=\underset{\sim}{\mu}}$. Then the limiting distribution is the same as the one of $\sqrt{n}\ \underset{\sim}{\xi}'(\underset{\sim}{x}_n - \underset{\sim}{\mu})$, which is $N(0,\underset{\sim}{\xi}'\Sigma\underset{\sim}{\xi})$.

Lemma 2.2. Let nS be distributed as a Wishart distribution $W_p(\Sigma,n)$ and B be distributed as a noncentral Wishart distribution $W_p(\Sigma,q:\Omega)$. Assume $\Omega = 0(n) = n\theta$. Then the limiting distributions of $V = \sqrt{n}\ (S - \Sigma)$ and $U = \sqrt{n}\ (\frac{1}{n} B - \theta)$ are multivariate normal. Further the limiting distributions of trCV and trCU are $N(0,\sigma_1^2)$ and $N(0,\sigma_2^2)$ respectively, where C is a symmetric matrix of order $p \times p$, $\sigma_1^2 = 2\ \mathrm{tr}(C\Sigma)^2$ and $\sigma_2^2 = 4\ \mathrm{tr}\ C^2\theta$.

This lemma is well known and is proved by considering the ch. functions of V and U.

Lemma 2.3. Let $S_{1,n}$ and $S_{2,n}$ be sequences of symmetric matrices of order $p \times p$ such that the limiting distributions of $V_{1,n} = \sqrt{n}\ (S_{1,n} - \Lambda)$ and $V_{2,n} = \sqrt{n}\ (S_{2,n} - I_p)$ are multivariate normal, where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_p)$, $\lambda_1 \geq,\ldots,\geq \lambda_p$ and I_p is the identity matrix of order $p \times p$. Let $\ell_1 \geq,\ldots,\geq \ell_p$ and $d_1 \geq,\ldots,\geq d_p$ be the eigenvalues of $S_{1,n}$ and $S_{1,n}S_{2,n}^{-1}$, respectively. Suppose that the α-th largest eigenvalue λ_α of Λ is simple. Then the limiting distributions of $\sqrt{n}\ (\ell_\alpha - \lambda_\alpha)$ and $\sqrt{n}\ (d_\alpha - \lambda_\alpha)$ are the same as the ones of $(V_{1,n})_{\alpha\alpha}$ and $(V_{1,n})_{\alpha\alpha} - (V_{2,n})_{\alpha\alpha}$ respectively, where $(A)_{\alpha\beta}$ denotes the (α,β)-th element of a matrix A.

This lemma has been essentially proved in the papers of Hsu (1941a,b) and Anderson (1963) who treated the general case of multiple roots.

3. PRINCIPAL COMPONENT ANALYSIS

In the area of principal component analysis, it is of interest to find out a small number of principal components which would adequately explain the variation among experimental units. In the population, the variance of i-th important principal component is the i-th largest eigenvalue of the population covariance matrix. If these eigenvalues are small, then the corresponding principal components are unimportant. In a number of situations, it is of interest to find out as to whether the addition of some variables will increase the variances of the first few important principal components. Similarly, it is of interest to find out whether there is significant increase in the ratio of the i-th largest eigenvalue to the trace of the covariance matrix if some variables are added. So, we will derive asymptotic distributions of increases in certain functions of the eigenvalues of the sample covariance matrix when a new set of variables is added.

Let $\underset{\sim}{x}: p \times 1$ be distributed as $N(\underset{\sim}{0}, \Sigma)$. We partition $\underset{\sim}{x} = (\underset{\sim}{x}_1', \underset{\sim}{x}_2')$, $\underset{\sim}{x}_1: p_1 \times 1$ and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \Sigma_{11}: p_1 \times p_1. \tag{3.1}$$

Suppose the vector $\underset{\sim}{x}_1$ is augmented to $\underset{\sim}{x}$. Let λ_α and $\tilde{\lambda}_\alpha$ be the α-th largest roots of Σ_{11} and Σ, respectively. Then

$$\tilde{\lambda}_\alpha \geq \lambda_\alpha \quad (\alpha = 1, \ldots, p_1)$$

which follows from the Poincaré separation theorem (see, e.g., Rao (1973, p.64)).

We are interested in the increases $\delta_\alpha = \lambda_\alpha - \lambda_\alpha$. Let S be the sample covariance matrix based on a sample of size $N = N + 1$. We partition S as in (3.1

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}. \tag{3.2}$$

The sample quantities corresponding to δ_α are

$$d_\alpha = \tilde{\ell}_\alpha - \ell_\alpha, \quad (\alpha = 1, \ldots, p_1)$$

where ℓ_α and $\tilde{\ell}_\alpha$ are the α-th largest roots of S_{11} and S, respectively. We consider the distribution of

$$J = \sqrt{n}\ \{f(d_1,\ldots,d_{p_1}) - f(\delta_1,\ldots,\delta_{p_1})\} \tag{3.3}$$

We assume

A1: $f(\underset{\sim}{d})$ is continuously differentiable in a neighborhood of $\underset{\sim}{d} = \underset{\sim}{\delta}$

where $\underset{\sim}{d} = (d_1,\ldots,d_{p_1})'$ and $\underset{\sim}{\delta} = (\delta_1,\ldots,\delta_{p_1})'$. Let

$$\underset{\sim}{c} = (c_1,\ldots,c_{p_1})' = \frac{\partial}{\partial \underset{\sim}{d}}\ f(\underset{\sim}{d})\Big|_{\underset{\sim}{d} = \underset{\sim}{\delta}} \tag{3.4}$$

Let $H_{11}: p_1 \times p_1$ be an orthogonal matrix such that

$$H_{11}\Sigma_{11}H_{11}' = \Lambda_{11} = \text{diag}\ (\lambda_1,\ldots,\lambda_{p_1}).$$

Since ℓ_α and $\tilde{\ell}_\alpha$ are invariant under the transformation

$$\underset{\sim}{x} \longrightarrow \begin{bmatrix} H_{11} & 0 \\ 0 & I \end{bmatrix} \underset{\sim}{x}.$$

we may assume

$$\Sigma = \Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \tag{3.5}$$

where $\Lambda_{12} = H_{11}\Sigma_{12}$, $\Lambda_{21} = \Sigma_{21}H_{11}'$ and $\Lambda_{22} = \Sigma_{22}$. Let

$$S = \Lambda + \frac{1}{\sqrt{n}}\, V \tag{3.6}$$

and

$$\tilde{S} = \Gamma' S \Gamma = \tilde{\Lambda} + \frac{1}{\sqrt{n}}\, \Gamma' V \Gamma \tag{3.7}$$

where Γ is an orthogonal matrix such that $\Gamma'\Lambda\Gamma = \tilde{\Lambda} = \text{diag}\,(\tilde{\lambda}_1,\ldots,\tilde{\lambda}_p)$. Since ℓ_α and $\tilde{\ell}_\alpha$ are the α-th largest roots of $S_{11} = \Lambda_{11} + (1/\sqrt{n})\, V_{11}$ and $\tilde{S} = \tilde{\Lambda} + (1/\sqrt{n})\, \Gamma'V\Gamma$, we obtain by Lemma 2.3 that the asymptotic distribution of $\sqrt{n}\,(d_\alpha - \delta_\alpha)$ is the same as that of

$$g_\alpha = (\Gamma'V\Gamma)_{\alpha\alpha} - (V)_{\alpha\alpha} \tag{3.8}$$

if λ_α and $\tilde{\lambda}_\alpha$ are simple. Using Lemma 2.1 we obtain that the asymptotic distribution of J is the same as that of

$$\sum_{\alpha=1}^{p_1} c_\alpha g_\alpha = \text{tr}\, A\, V \tag{3.9}$$

where

$$A = \Gamma D_c \Gamma' - D_c,\quad D_c = \text{diag}(c_1,\ldots,c_{p_1},0,\ldots,0). \tag{3.10}$$

This implies the following.

Theorem 3.1. Let nS be distributed as a Wishart distribution $W_p(\Sigma,n)$. Let $d_\alpha = \tilde{\ell}_\alpha - \ell_\alpha$ and $\delta_\alpha = \tilde{\lambda}_\alpha - \lambda_\alpha$, where ℓ_α, $\tilde{\ell}_\alpha$, λ_α and $\tilde{\lambda}_\alpha$ are the α-th largest roots of S_{11}, S, Σ_{11} and Σ. Assume a function $f(d_1,\ldots,d_{p_1})$ satisfies the assumption A1 and all the roots λ_α and $\tilde{\lambda}_\alpha$ $(\alpha = 1,\ldots,p_1)$ are simple. Then

$$\sqrt{n}\ \{f(d_1,\ldots,d_{p_1}) - f(\delta_1,\ldots,\delta_{p_1})\} \xrightarrow{D} N(0,\sigma^2)$$

as $n \to \infty$, where

$$\begin{aligned}\sigma^2 &= 2\ \text{tr}\ (A\Lambda)^2 \\ &= 2 \sum_{\alpha=1}^{p_1} c_\alpha^2(\lambda_\alpha^2 + \tilde{\lambda}_\alpha^2) - 4 \sum_{\alpha,\beta=1}^{p_1} c_\alpha c_\beta \tilde{\lambda}_\alpha^2 \gamma_{\beta\alpha}^2\end{aligned}$$

and $\Gamma = (\gamma_{\alpha\beta})$.

Cor. 3.1.1. When λ_α and $\tilde{\lambda}_\alpha$ are simple,

$$\sqrt{n}\ \{(\tilde{\ell}_\alpha - \ell_\alpha) - (\tilde{\lambda}_\alpha - \lambda_\alpha)\} \xrightarrow{D} N(0,\sigma^2)$$

as $n \to \infty$, where $\sigma^2 = 2(\lambda^2 + \tilde{\lambda}^2) - 4\,\tilde{\lambda}_\alpha^2 \gamma_{\alpha\alpha}^2$.

4. EFFECT OF ADDITIONAL VARIABLES IN DISCRIMINANT ANALYSIS

In the area of discriminant analysis, it is of interest to find out as to whether the addition of a new set of variables will make a significant contribution on the discriminant functions. This problem can be investigated by examining the increases due to the additional variables in certain functions of the eigenvalues of the MANOVA matrix. So, we will study the asymptotic distributions of the above increases in the sample.

Let W and B be independently distributed as a Wishart distribution $W_p(\Sigma,n)$ and a noncentral Wishart distribution $W_p(\Sigma,q:\Xi)$

We partition

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \qquad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \Xi = \begin{pmatrix} \Xi_{11} & \Xi_{12} \\ \Xi_{21} & \Xi_{22} \end{pmatrix}, \tag{4.1}$$

with $W_{11}: p_1 \times p_1$, $B_{11}: p_1 \times p_1$, $\Sigma_{11}: p_1 \times p_1$ and $\Xi_{11}: p_1 \times p_1$. Let ℓ_α, $\tilde{\ell}_\alpha$, ω_α and $\tilde{\omega}_\alpha$ be the α-th largest roots of $B_{11}W_{11}^{-1}$, BW^{-1}, $\Xi_{11}\Sigma_{11}^{-1}$ and $\Xi\Sigma^{-1}$, respectively. Then

$$\begin{aligned} d_\alpha &= \tilde{\ell}_\alpha - \ell_\alpha \geq 0, \\ n\delta_\alpha &= \tilde{\omega}_\alpha - \omega_\alpha \geq 0, \ (\alpha = 1,\ldots,p_1) \end{aligned} \tag{4.2}$$

which follows from the Poincaré separation theorem in the case of two matrices (also see Gabriel (1968)).

Let

$$T = H\,L = \begin{pmatrix} H_{11} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \tag{4.3}$$

where $L\Sigma L' = I$ and H_{11} is an orthogonal matrix such that $H_{11}L_{11}\Xi_{11}L_{11}'H_{11}' = \Omega_{11} = \mathrm{diag}(\omega_1,\ldots,\omega_{p_1})$. Since ℓ_α and $\tilde{\ell}_\alpha$ are invariant under the transformation $B \longrightarrow TBT'$ and $W \longrightarrow TWT'$, we may assume

$$\Sigma = I,$$

$$\Xi = \Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \tag{4.4}$$

with $\Omega_{11} = \mathrm{diag}(\omega_1,\ldots,\omega_{p_1})$, $\Omega_{12} = H_{11}L_{11}(\Xi_{11}L_{21}' + \Xi_{12}L_{22}')$, $\Omega_{21} = \Omega_{12}'$ and $\Omega_{22} = (L_{21}\Xi_{11} + L_{22}\Xi_{21})L_{21}' + (L_{21}\Xi_{12} + L_{22}\Xi_{22})L_{22}'$. We assume

$$\text{A2:}\quad \Omega = O(n)$$

$$= n\,\Theta = n\begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{pmatrix} \tag{4.5}$$

where $\Theta_{11} = \mathrm{diag}(\theta_1,\ldots,\theta_{p_1})$. Let Γ be an orthogonal matrix such that

$$\Gamma'\,\Theta\,\Gamma = \tilde{\Theta} = \mathrm{diag}(\tilde{\theta}_1,\ldots,\tilde{\theta}_p) \tag{4.6}$$

with $\tilde{\theta}_1 \geq \ldots \geq \tilde{\theta}_p$. Then $n\tilde{\theta}_\alpha = \tilde{\omega}_\alpha$. Let

$$\frac{1}{n}B = \Theta + \frac{1}{\sqrt{n}}U,$$

$$\frac{1}{n}W = I + \frac{1}{\sqrt{n}}W. \tag{4.7}$$

Then it is easily seen that ℓ_α and $\tilde{\ell}_\alpha$ are the α-th largest roots of

$$\left|\Theta_{11} + \frac{1}{\sqrt{n}}U_{11} - \ell\left(I - \frac{1}{\sqrt{n}}V_{11}\right)\right| = 0$$

and

$$\left|\tilde{\Theta} + \frac{1}{\sqrt{n}}\Gamma'U\Gamma - \tilde{\ell}\left(I + \frac{1}{\sqrt{n}}\Gamma'V\Gamma\right)\right| = 0$$

respectively, where $U_{11}: p_1 \times p_1$ and $V_{11}: p_1 \times p_2$ are the submatrices of U and V partitioned as in (4.1).

From Lemma 2.3, it is seen that the asymptotic distribution of $\sqrt{n}\ (d_\alpha - \delta_\alpha)$ is the same as that of

$$g_\alpha = (\Gamma'U\Gamma)_{\alpha\alpha} - \tilde{\theta}_\alpha(\Gamma'V\Gamma)_{\alpha\alpha} - (V)_{\alpha\alpha} + \theta_\alpha(V)_{\alpha\alpha} \tag{4.8}$$

if ω_α and $\tilde{\omega}_\alpha$ are simple. Using Lemma 1, we obtain that the asymptomatic distribution of $\sqrt{n}\ \{f(d_1,\ldots,d_{p_1}) - f(\delta_1,\ldots,\delta_{p_1})\}$ is the same as that of

$$\sum_{\alpha=1}^{p_1} c_\alpha g_\alpha = \mathrm{tr}A^{(1)}V + \mathrm{tr}A^{(2)}U \tag{4.9}$$

where $A^{(1)} = D_{c\theta} - D_c\tilde{\theta}\Gamma'$, $A^{(2)} = \Gamma D_c\Gamma' - D_c$, $D_c = \mathrm{diag}(c_1,\ldots,c_{p_1},0,\ldots,0)$, $D_{c\theta} = \mathrm{diag}(c_1\theta_1,\ldots,c_{p_1},0,\ldots,0)$, etc. This implies the following.

Theorem 4.1. Let W and B be independently distributed as a Wishart distribution $W_p(\Sigma,n)$ and a noncentral Wishart distribution $W_p(\Sigma,q:\Xi)$, respectively. Let $d_\alpha = \tilde{\ell}_\alpha - \ell_\alpha$ and $n\delta_\alpha = \tilde{\omega}_\alpha - \omega_\alpha$, where ℓ_α, $\tilde{\ell}_\alpha$, ω_α and $\tilde{\omega}_\alpha$ are the α-th largest roots of $B_{11}W_{11}^{-1}$, BW^{-1}, $\Xi_{11}\Sigma_{11}^{-1}$ and $\Xi\Sigma^{-1}$, respectively. Assume that the assumptions A1 and A2 are satisfied, and ω_α and $\tilde{\omega}_\alpha$ $(\alpha = 1,\ldots,p_1)$ are simple. Then

$$\sqrt{n}\ \{f(d_1,\ldots,d_{p_1}) - f(\delta_1,\ldots,\delta_{p_1})\} \xrightarrow{D} N(0,\sigma^2)$$

as $n \to \infty$, where

$$\begin{aligned}\sigma^2 &= 2\ \mathrm{tr}\{D_{c\theta} - \Gamma D_c\tilde{\theta}\Gamma'\}^2 + 4\ \mathrm{tr}\{\Gamma D_c\Gamma' - D_c\}^2\Theta \\ &= 2\sum_{\alpha=1}^{p_1} c_\alpha^2\ \{\theta_\alpha(\theta_\alpha + 2) + \tilde{\theta}_\alpha(\tilde{\theta}_\alpha + 2) \\ &\quad - 4\sum_{\alpha,\beta=1}^{p_1} c_\alpha c_\beta\tilde{\theta}_\beta\gamma_{\alpha\beta}^2\ (\theta_\alpha + 2)\end{aligned}$$

and $\Gamma = (\gamma_{\alpha\beta})$.

Cor. 4.1.1. When ω_α and $\tilde{\omega}_\alpha$ are simple

$$\sqrt{n}\,[(\tilde{\ell}_\alpha - \ell_\alpha) - (\tilde{\theta}_\alpha - \theta_\alpha)] \xrightarrow{D} N(0,\sigma^2)$$

as $n \to \infty$, where $\sigma^2 = 2\,\theta_\alpha(\theta_\alpha + 2) + 2\,\tilde{\theta}_\alpha(\tilde{\theta}_\alpha + 2) - 4\,\tilde{\theta}_\alpha(\theta_\alpha + 2)\gamma^2_{\alpha\alpha}$.

5. EFFECT OF ADDITIONAL VARIABLES ON CANONICAL VARIABLES

In this section, we study asymptotic distributions of certain statistics useful in studying the effect of additional variables on the canonical correlations.

Consider two sets of variables $\underset{\sim}{x}_1$: $p_1 \times 1$ and $\underset{\sim}{y}_1$: $q_1 \times 1$. We assume $p_1 \leq q_1$. Let $\rho_1 \geq \ldots \geq \rho_{p_1} \geq 0$ be the canonical correlations between $\underset{\sim}{x}_1$ and $\underset{\sim}{y}_1$. We shall augment the variates $\underset{\sim}{x}_1$ and $\underset{\sim}{y}_1$ to $\underset{\sim}{x}$: $p \times 1$ and $\underset{\sim}{y}$: $q \times 1$ by adding extra variates $\underset{\sim}{x}_2$: $p_2 \times 1$ and $\underset{\sim}{y}_2$: $q_2 \times 1$, respectively. We assume that $(\underset{\sim}{x}', \underset{\sim}{y}')$ is distributed as $N_{p+q}[\mu, \Sigma]$. We partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}, \quad \Sigma_{xx}: p \times p. \tag{5.1}$$

Let $\tilde{\rho}_\alpha$ be the α-th largest canonical correlation between $\underset{\sim}{x}$ and $\underset{\sim}{y}$. Then

$$\delta_\alpha = \tilde{\rho}_\alpha - \rho_\alpha \geq 0 \quad (\alpha = 1, \ldots, p_1). \tag{5.2}$$

which has been shown in Fujikoshi (1982).

Let

$$S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix}$$

be the sample covariance matrix besed on a sample of size $N = n + 1$. Let r_α and r_α be the α-th largest canonical correlations between $\underset{\sim}{x}_1$ and $\underset{\sim}{y}_1$ and between $\underset{\sim}{x}$ and $\underset{\sim}{y}$. We consider the asymptotic distribution of

$$\sqrt{n}\,\{f(d_1, \ldots, d_{p_1}) - f(\delta_1, \ldots, \delta_{p_1})\}$$

where $d_\alpha = \tilde{r}_\alpha - r_\alpha$. Let L_1 and L_2 be the lower triangular matrices such that

$$L_1 \Sigma_{xx} L_1' = I_p, \quad L_2 \Sigma_{yy} L_2' = I_q. \tag{5.3}$$

We partition

$$L_1 = \begin{pmatrix} L_{1\cdot 11} & 0 \\ L_{1\cdot 21} & L_{1\cdot 22} \end{pmatrix}, \quad L_2 = \begin{pmatrix} L_{2\cdot 11} & 0 \\ L_{2\cdot 21} & L_{2\cdot 22} \end{pmatrix}. \tag{5.4}$$

Let $H_{1\cdot 11}$: $p_1 \times p_1$ and $H_{2\cdot 11}$: $p_2 \times p_2$ be the orthogonal matrices chosen so that

$$H_{1\cdot 11} L_{1\cdot 11}\, \Sigma_{xy\cdot 11} L'_{2\cdot 11} H'_{2\cdot 11} = P_{11} = \begin{pmatrix} \rho_1 & & 0 \\ & \ddots & \\ 0 & & \rho_{p_1} \end{pmatrix}, \tag{5.5}$$

where

$$\Sigma_{xy} = \begin{pmatrix} \Sigma_{xy\cdot 11} & \Sigma_{xy\cdot 12} \\ \Sigma_{xy\cdot 21} & \Sigma_{xy\cdot 22} \end{pmatrix}, \ \Sigma_{xy\cdot 11}: p_1 \times q_1. \tag{5.6}$$

Then r_α and $\tilde{r}_\alpha$ are invariant under the transformation

$$\underset{\sim}{x} \longrightarrow \begin{pmatrix} H_{1\cdot 11} L_{1\cdot 11} & 0 \\ L_{1\cdot 21} & L_{1\cdot 22} \end{pmatrix} \underset{\sim}{x},$$

$$\underset{\sim}{y} \longrightarrow \begin{pmatrix} H_{2\cdot 11} L_{2\cdot 11} & 0 \\ L_{2\cdot 21} & L_{2\cdot 22} \end{pmatrix} \underset{\sim}{y}.$$

Therefore, we may assume

$$\Sigma_{xx} = I_p, \ \Sigma_{yy} = I_q$$

$$\Sigma_{xy} = P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \tag{5.7}$$

with P_{11} in (5.5), $P_{12} = H_{1\cdot 11}L_{1\cdot 11}(\Sigma_{11}L'_{2\cdot 21} + \Sigma_{12}L'_{2\cdot 22})$, $P_{21} = (L_{1\cdot 21}\Sigma_{11} + L_{1\cdot 22}\Sigma_{21})L'_{2\cdot 11}H'_{2\cdot 11}$ and $P_{22} = (L_{1\cdot 21}\Sigma_{11} + L_{1\cdot 22}\Sigma_{21})L'_{2\cdot 21} + (L_{1\cdot 21}\Sigma_{12} + L_{1\cdot 22}\Sigma_{22})L'_{2\cdot 22}$. Let

$$S = \Sigma + \frac{1}{\sqrt{n}} V$$

$$= \begin{bmatrix} I_p & P \\ P' & I_q \end{bmatrix} + \frac{1}{\sqrt{n}} \begin{bmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{bmatrix} . \qquad (5.8)$$

Then r_α is the α-th root of

$$|S_{xy\cdot 11}S^{-1}_{yy\cdot 11}S_{yx\cdot 11} - r\, S_{xx\cdot 11}| = 0$$

which is equivalent to

$$P_{11}P'_{11} + \frac{1}{\sqrt{n}} Z_{11} - r^2(I_{p_1} + \frac{1}{\sqrt{n}} V_{xx\cdot 11})| = 0 \qquad (5.9)$$

where

$$Z_{11} = \sqrt{n}\{(P_{11} + \frac{1}{\sqrt{n}} V_{xy\cdot 11})(I_{q_1} + \frac{1}{\sqrt{n}} V_{y\cdot 11})^{-1}(P'_{11} + \frac{1}{\sqrt{n}} V_{yx\cdot 11}) - P_{11}P'_{11}\}, \qquad (5.10)$$

$S_{xy\cdot 11}$, $V_{xy\cdot 11}$, etc. denote the submatrices of S_{xy} and V_{xy} partitioned as in (5.6). Let Γ_1 and Γ_2 be the orthogonal matrices such that

$$\Gamma_1 P \Gamma'_2 = \tilde{P}: \; p \times q \qquad (5.11)$$

where the (α,α) elements of $\tilde{P}$ are $\tilde{\rho}_\alpha$ and other elements are zero. Let

$$\tilde{S} = \begin{bmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{bmatrix} S \begin{bmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{bmatrix} = \begin{bmatrix} \Gamma_1 S_{xx}\Gamma'_1 & \Gamma_1 S_{xy}\Gamma'_2 \\ \Gamma_2 S_{yx}\Gamma'_2 & \Gamma_2 S_{yy}\Gamma'_2 \end{bmatrix}$$

$$= \begin{bmatrix} I_p & P \\ P' & I_q \end{bmatrix} + \frac{1}{\sqrt{n}} \begin{bmatrix} \Gamma_1 V_{xx}\Gamma'_1 & \Gamma_1 V_{xy}\Gamma'_2 \\ \Gamma_2 V_{yx}\Gamma'_1 & \Gamma_2 V_{yy}\Gamma'_2 \end{bmatrix} . \qquad (5.12)$$

From (5.12) we obtain that r_α is the α-th largest root of

$$\left| \tilde{P}\tilde{P}' + \frac{1}{\sqrt{n}} \tilde{Z} - \tilde{r}^2 (I + \frac{1}{\sqrt{n}} \Gamma_1 V_{xx} \Gamma_1') \right| = 0. \tag{5.13}$$

where

$$\tilde{Z} = \sqrt{n}\{\tilde{P} + \frac{1}{\sqrt{n}} \Gamma_1 V_{xx} \Gamma_1')(I_q + \frac{1}{\sqrt{n}} \Gamma_2 V_{yy} \Gamma_2')^{-1} (\tilde{P} + \frac{1}{\sqrt{n}} \Gamma_2 V_{yx} \Gamma_1') - \tilde{P}\tilde{P}'\} \tag{5.14}$$

We note that the limiting distributions of Z_{11} and $\tilde{Z}$ are the same as those of $V_{xy \cdot 11} P_{11}' + P_{11} V_{yx \cdot 11} - P_{11} V_{yy \cdot 11} P_{11}'$ and $\Gamma_1 V_{xy} \Gamma_1' \tilde{P}' + \tilde{P} \Gamma_2 V_{yx} \Gamma_1'$ $- \tilde{P} \Gamma_2 V_{yy} \Gamma_2' \tilde{P}'$, respectively. Applying Lemma 2.3 to (5.9) and (5.14) we obtain that the asymptotic distribution of $\sqrt{n}\ (d_\alpha - \delta_\alpha)$ is the same as that of

$$\begin{aligned} g_\alpha = {} & (\Gamma_1 V_{xy} \Gamma_2')_{\alpha\alpha} - (V_{xy})_{\alpha\alpha} \\ & - \frac{1}{2} \tilde{\rho}_\alpha\ (\Gamma_1 V_{xx} \Gamma_1')_{\alpha\alpha} - \frac{1}{2} \tilde{\rho}_\alpha\ (\Gamma_2 V_{yy} \Gamma_2')_{\alpha\alpha} \\ & + \frac{1}{2} \tilde{\rho}_\alpha\ (V_{xx})_{\alpha\alpha} + \frac{1}{2}\ \rho_\alpha\ (V_{yy})_{\alpha\alpha}. \end{aligned} \tag{5.15}$$

Therefore the asymptotic distribution of $\sqrt{n}\ \{f(d_1,\ldots,d_{p_1}) - f(\delta_1,\ldots,\delta_{p_1})\}$ is the same as that of

$$\sum c_\alpha g_\alpha = \frac{1}{2} \operatorname{tr} AV \tag{5.16}$$

where

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

$$A_{11} = \operatorname{diag}(c_1\rho_1,\ldots,c_{p_1}\rho_{p_1},0,\ldots,0) - \Gamma_1' \operatorname{diag}\ (\tilde{c}_1\rho_1,\ldots,c_{p_1}\tilde{\rho}_{p_1},0,\ldots,0)\Gamma_1,$$

$$A_{22} = \operatorname{diag}(c_1\rho_1,\ldots,c_{p_1}\rho_{p_1},0,\ldots,0) - \Gamma_2' \operatorname{diag}(c_1\tilde{\rho}_1,\ldots,c_{p_1}\tilde{\rho}_{p_1},0,\ldots,0)\Gamma_2,$$

$$A_{21} = A_{12}' = \Gamma_2' D_c\ {}_1\Gamma - D_c$$

and the (α,α) elements of D_c: $q \times p$ are c_α for $\alpha = 1,\ldots,p_1$ and the other elements of D_c are zero. This implies the following.

Theorem 5.1. Let r_α and $\tilde{r}_\alpha$ be the α-th largest canonical correlations between $\underset{\sim}{x}_1$: $p_1 \times 1$ and $\underset{\sim}{y}_1$: $q_1 \times 1$ $(p_1 \le q_1)$ and between $\underset{\sim}{x}$: $p \times 1$ and $\underset{\sim}{y}$: $q \times 1$, based on a sample of size $N = n+1$ from $N(\underset{\sim}{\mu}, \Sigma)$.

Let ρ_α and $\tilde{\rho}_\alpha$ be the corresponding population quantities. Assume that a function $f(d_1,\ldots,d_{p_1})$ satisfies the assumption A1 and the canonical correlations ρ_α and ρ_α $(\alpha = 1,\ldots,p_1)$ are simple. Then

$$\sqrt{n}\,\{f(d_1,\ldots,d_{p_1}) - f(\delta_1,\ldots,\delta_{p_1})\} \xrightarrow{D} N(0,\sigma^2)$$

as $n \to \infty$, where $d_\alpha = \tilde{r}_\alpha - r_\alpha$, $\delta_\alpha = \tilde{\rho}_\alpha - \rho_\alpha$,

$$\sigma^2 = \frac{1}{2}\,\mathrm{tr}\left\{\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} I_p & P \\ P' & I_q \end{pmatrix}\right\}^2$$

$$= \sum_{\alpha,\beta=1}^{p_1} c_\alpha^2\,\{(1-\rho_\alpha^2)^2 + (1-\tilde{\rho}_\alpha^2)^2\}$$

$$+ \sum_{\alpha,\beta=1}^{p_1} c_\alpha c_\beta\,(1-\tilde{\rho}_\beta^2)\,\{\rho_\alpha\tilde{\rho}_\beta(\gamma_{1\cdot\beta\alpha}^2 + \gamma_{2\cdot\beta\alpha}^2) - 2\gamma_{1\cdot\beta\alpha}\gamma_{2\cdot\beta\alpha}\}$$

and $\Gamma_1 = (\gamma_{1\cdot\alpha\beta})$ and $\Gamma_2 = (\gamma_{2\cdot\alpha\beta})$.

Cor. 5.1.1. When ρ_α and $\tilde{\rho}_\alpha$ are simple,

$$\sqrt{n}\,\{(\tilde{r}_\alpha - r_\alpha) - (\tilde{\rho}_\alpha - \rho_\alpha)\} \xrightarrow{D} N(0,\sigma^2)$$

as $n \to \infty$, where $\sigma^2 = (1-\tilde{\rho}_\alpha^2)^2 + (1-\tilde{\rho}_\alpha^2)[\rho_\alpha\tilde{\rho}_\alpha(\gamma_{1\cdot\alpha\alpha}^2 + \gamma_{2\cdot\alpha\alpha}^2) - 2\gamma_{1\cdot\alpha\alpha}\gamma_{2\cdot\alpha\alpha}]$

Note: Wijsman (1984) proved the Corr. 5.1.1 in the case of $\alpha = 1$ and $q_2 = 0$ or $p_2 = 0$.

6. FURTHER RESULTS

Let d_α and δ_α be the increases in the α-th largest sample and population eigenvalues in the three cases: (i) principal component analysis, (ii) discriminant analysis, and (iii) canonical correlation analysis. Then in Theorems 3.1, 4.1, and 5.1, we have shown that

$$J = \sqrt{n}\,\{f(d_1,\ldots,d_{p_1}) - f(\sigma_1,\ldots,\sigma_{p_1})\} \xrightarrow{D} N(0,\sigma^2). \tag{6.1}$$

The limiting variance σ^2 depends on unknown Σ for cases (i) and (iii) and unknown Σ and Θ for case (ii). In order to make the formula useful, we need to estimate $\hat{\sigma}^2$ obtained from σ^2 by replacing Σ by S for (i) and (iii), and by replacing Σ and Θ by $(1/n)W$ and $(1/n)B$, respectively, for (ii). It is easy to see that under the same assumption as in each of the Theorems

$$\hat{\sigma} \to \delta \quad \text{in probability.}$$

Therefore, it follows from (6.1) that

$$\hat{\sigma}^{-1} J \xrightarrow{D} N(0,1), \tag{6.2}$$

the formula is useful in constructing an approximate confidence interval for $f(\delta_1,\ldots,\delta_{p_1})$.

It is easy to extend the result for a single function J to the one for several functions. Let

$$J_\alpha = \sqrt{n}\,\{f_\alpha(d_1,\ldots,d_{p_1}) - f_\alpha(\delta_1,\ldots,\delta_{p_1})\}$$

for $\alpha = 1,2,\ldots,k$. We assume that f_α's satisfy the assumption A1. Let

$$\underset{\sim}{c}_\alpha = (c_{1,\alpha},\ldots,c_{p_1,\alpha})' = \frac{\partial}{\partial \underset{\sim}{d}} f_\alpha(\underset{\sim}{d}) \Big|_{\underset{\sim}{d} = \underset{\sim}{\delta}}.$$

Then we can prove that

$$\underset{\sim}{J} = (J_1,\ldots,J_k) \xrightarrow{D} N_k(\underset{\sim}{0},Q). \tag{6.3}$$

The limiting covariance matrix $Q = (q_{\alpha\beta})$ is given as follows:

Case (i):

$$q_{\alpha\beta} = 2\ \mathrm{tr}(A_\alpha \Lambda A_\beta \Lambda),$$

where A_α is defined from A in (3.10) by substituting $\underset{\sim}{c}$ into $\underset{\sim}{c}_\alpha$.

Case (ii)

$$q_{\alpha\beta} = 2\ \mathrm{tr}A_\alpha^{(1)}A_\beta^{(1)} + 4\ \mathrm{tr}A_\alpha^{(2)}A_\beta^{(2)}\Theta,$$

where $A_\alpha^{(1)}$ and $A_\alpha^{(2)}$ are defined from the $A^{(1)}$ and $A^{(2)}$ in (4.9) by substituting $\underset{\sim}{c}$ into $\underset{\sim}{c}_\alpha$.

Case (iii)

$$q_{\alpha\beta} = \frac{1}{4}\ \mathrm{tr}A_\alpha \begin{bmatrix} I_p & P \\ P' & I_q \end{bmatrix} A_\beta \begin{bmatrix} I_p & P \\ P' & I_q \end{bmatrix},$$

where A_α is defined from A in (5.14) by substituting $\underset{\sim}{c}$ into $\underset{\sim}{c}_\alpha$.

Higher order terms of the joint distribution of $J_1,\ldots,J_k$ can be obtained by using perturbation technique.

REFERENCES

[1] Anderson, T.W. (1963). Asymptotic theory for principal component analysis. Ann. Math. Statist. 34, 122-148.

[2] Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

[3] Fujikoshi, Y. (1982). A test for additional information in canonical correlation analysis. Ann. Inst. Statist. Math. 34, 523-530.

[4] Gabriel, K.R. (1968). Simultaneous test procedures in multivariate analysis of variance. Biometrika 55, 484-504.

[5] Hsu, P.L. (1941a). On the limiting distribution of roots of a determinantal equation. J. London Math. Soc. 16, 183-194.

[6] Hsu, P.L. (1941b). On the limiting distribution of the canonical correlations. Biometrika 32, 38-45.

[7] Rao, C.R. (1966). Covariance adjustment and related problems in multivariate analysis. In Multivariate Analysis (P.R. Krishnaiah, editor), 87-103. Academic Press, New York.

[8] Rao, C.R. (1973). Linear Statistical Inference and Its Applications. New York, John Wiley & Sons.

[9] Wijsman, R.A. (1984). Asymptotic distribution of the increase of the largest canonical correlation when one of the vectors is augmented. To appear in J. Multivariate Anal.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 90	2 GOVT ACCESSION NO. AD-A162 069	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Effect of Additional Variables in Principal Component Analysis, Discriminant Analysis and Canonical Correlation Analysis		5 TYPE OF REPORT & PERIOD COVERED Technical - August, 1985
		6. PERFORMING ORG. REPORT NUMBER 85-31
7 AUTHOR(s) Y.Fujikoshi, P.R.Krishnaiah, J.Schmidhammer		8. CONTRACT OR GRANT NUMBER(s) F49620-85-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Multivariate Analysis 515 Thackeray Hall, University of Pittsburgh Pittsburgh, Pennsylvania 15260		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F, 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Department of the Air Force Bolling Air Force Base, DC 20332		12. REPORT DATE August 1985
		13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Asymptotic distribution theory, Canonical correlation analysis, Discriminant analysis, Principal component analysis.

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this paper, the authors derived asymptotic distributions of changes in certain functions of the eigenvalues of the sample covariance matrix, MANOVA matrix and canonical correlation matrix when some variables are added to the original sets of variables. The above results are useful in finding out as to whether the new variables give additional information for Statistical inference in multivariate analysis, Wishart distribution.

DD FORM 1 JAN 73 1473

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

END

FILMED

1-86

DTIC